Bilingual movie sentiment analysis

Rafik Youssef, Seif El-Berkawy

May 7, 2023

Abstract

In this paper, we present a deep learning-based sentiment analysis system for Arabic and English movie reviews. To begin, we preprocess the reviews using stemming, stop word removal, and tokenization. On the preprocessed data, we train a Long Short-Term Memory (LSTM) neural network to forecast the sentiment of a specific review. Our algorithm successfully classified movie reviews as positive or negative with an accuracy of over 85

1 Introduction

1.1 Motivation

The aim of this project is to develop a sentiment analysis model for movie reviews in both Arabic and English. Sentiment analysis is a key phase in the processing of natural language that can offer useful knowledge about people's attitudes and opinions towards a given subject. Due to the growing popularity of social media platforms and the huge amount of user-generated stuff, sentiment analysis has increased in importance as a field of study for businesses of all sizes. Most of existing sentiment analysis algorithms, were created for English text and might not work as well with texts in other languages. By creating a model that can precisely identify Arabic and English movie reviews as good or negative, this study tackles this limitation. The suggested strategy can be utilized for many different kinds of purposes, including online review systems, market research and social media monitoring can give us important information about the tastes and opinions of our customers.

1.2 Challenges

While creating our sentiment analysis model for Arabic and English movie reviews, we encountered a number of difficulties. The lack of readily available datasets for Arabic movie reviews was one of the major difficulties. We spent three days looking for an Arabic movie reviews dataset because the majority of existing sentiment analysis datasets are in English. At first, we thought we would have to collect the data ourselves, but after extensive searching, we eventually located one. Additionally, the complexity of the morphology and the scarcity of language resources in Arabic present certain particular difficulties that made preprocessing and modelling more difficult. Additionally, the datasets contained reviews with varying lengths.

To solve these challenges, we preprocessed the data by deleting stop words, punctuation, and other unnecessary characters in order to get around these problems. Additionally, we employed padding to make sure that all reviews were the same length and a tokenizer that could accept Arabic text.

1.3 Datasets

In this project, we used two datasets for sentiment analysis of movie reviews in Arabic and English.

The first dataset is the "Arabic Movie Reviews Dataset", which is a publicly available, consisting of 50001 movie reviews in Arabic. The dataset is divided into two classes: positive and negative reviews, with 25001 positive sample and 25000 negative sample. The reviews were collected from various Arabic movie review websites, and they cover a wide range of movie genres.

The second dataset we used is the "IMDb Movie Reviews Dataset", which is a popular dataset for sentiment analysis of movie reviews in English. It contains 50,000 movie reviews, split into 25,000 for each class. The reviews are labeled as positive or negative, and they were collected from the IMDb website.

2 Data Analysis

The Data retrieved from IMDb did not need further cleaning, we were able to see that the reviews were evenly distributed on both positive and negative classes.

Both datasets were preprocessed to remove stop words, punctuation, and other unnecessary characters. In particular the English dataset was converted to lower case, punctuation and stop words removed. It was then tokenized and lemmatized using the NLTk Libary [BKL09]. On the other hand, the Arabic dataset was preprocessed by removing diacritics, stop words, punctuation and English characters and words afterwards tokenization and stemming, using ISRIStemmer [SKHD19], were applied. Finally, both datasets were concatenated together for the model to be trained on a single dataset.

References

[BKL09] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[SKHD19] Mochamad Gilang Syarief, Opik Taupik Kurahman, Arief Fatchul Huda, and Wahyudin Darmalaksana. Improving arabic stemmer: Isri stemmer. In *2019 IEEE 5th International Conference on Wireless and Telematics (ICWT)*, pages 1–4, 2019.